Understanding dynamic scenes based on human sequence evaluation ☆

Jordi Gonzàlez [a,*], Daniel Rowe [b], Javier Varona [c], F. Xavier Roca [b]

[a] *Institut de Robòtica i Informàtica Industrial (UPC-CSIC), Edifici U, Parc Tecnològic de Barcelona, Barcelona, Catalonia, Spain*
[b] *Computer Vision Center & Department of Computer Sciences (UAB), Edifici O, Campus UAB, Bellaterra, Catalonia, Spain*
[c] *Unitat de Gràfics i Visió per Ordinador (UIB), Edifici Anselm Turmeda, Campus UIB, Palma de Mallorca, Spain*

Abstract

In this paper, a Cognitive Vision System (CVS) is presented, which *explains* the human behaviour of monitored scenes using natural-language texts. This cognitive analysis of human movements recorded in image sequences is here referred to as Human Sequence Evaluation (HSE) which defines a set of transformation modules involved in the automatic generation of semantic descriptions from pixel values. In essence, the trajectories of human agents are obtained to generate textual interpretations of their motion, and also to infer the conceptual relationships of each agent w.r.t. its environment. For this purpose, a human behaviour model based on Situation Graph Trees (SGTs) is considered, which permits both bottom-up (hypothesis generation) and top-down (hypothesis refinement) analysis of dynamic scenes. The resulting system prototype interprets different kinds of behaviour and reports textual descriptions in multiple languages.

Keywords: Image Sequence Evaluation; High-level processing of monitored scenes; Segmentation and tracking in complex scenes; Event recognition in dynamic scenes; Human motion understanding; Human behaviour interpretation; Natural-language text generation; Realistic demonstrators

1. Introduction

During the past three decades, important research efforts in computer vision have been focused on developing theories, methods and systems applied to the description of human movements in image sequences. Broadly speaking, in the past the main goal was the estimation of quantitative parameters describing where was motion [11,1]. Nowadays, the focus is on the analysis of image sequences incorporating cognitive processes which allow to *understand* recorded movements [7,43,25]. That is, the true challenge is the generation of qualitative descriptions about the meaning of motion, therefore, understanding not only where, but also why motion is being observed. This has become a key task in many promising computer vision applications, such as smart video surveillance [33].

Towards this end, different Cognitive Vision Systems (CVS) have been proposed in the literature [42,28], which typically encompass topics not only related to computer vision, but also to artificial intelligence and computational linguistics. Here, we restrict cognition to assure the generation of plausible semantic interpretations of human movements observed in image sequences. In this context, the term *Human Sequence Evaluation* (HSE) denotes those transformation processes involved in the textual descriptions of human behaviour from pixel values [13]. Mainly, three co-operating goals are involved in HSE: (i) a geometric description of the recorded human movements is first obtained; (ii) subsequently, these quantitative parameters are used to instantiate logic predicates; and (iii) textual

☆ This work is supported by EC grants IST-027110 for the HERMES project and IST-045547 for the VIDI-video project, and by the Spanish MEC under projects TIN2006-14606 and CONSOLIDER-INGENIO 2010 (CSD2007-00018). Jordi Gonzàlez and Javier Varona also acknowledges the support of a Juan de la Cierva and a Ramon y Cajal Postdoctoral fellowships from the Spanish MEC, respectively.

* Corresponding author. Tel.: +34 935813841.
E-mail address: poal@cvc.uab.es (J. Gonzàlez).

representations are built upon these conceptual primitives for Natural-Language (NL) text generation.

Unfortunately, dealing with humans in image sequences implies several difficulties: a huge appearance variability is found due to acquisition conditions, clothes, lighting, and posture changes. Moreover, conceptual interpretations include uncertainties due to the vagueness of the semantic terms utilized. So multiple issues should be confronted for HSE, such as detection and localization; tracking; classification and categorization; prediction; concept formation; communication and expression; etc. Due to this complexity, HSE is designed as a mixture of co-operating top-down and bottom-up modules to define the interactions of computer vision algorithms with other components, such as human behaviour modeling and NL generation.

This contribution is structured as follows: we first identify the main tasks to be tackled for HSE by reviewing the literature. These transformation steps are embedded within an architecture based on co-operating modules. Subsequently, concrete example implementations of this architecture are presented for two scenes, i.e. a crosswalk and a cafeteria. We describe the main sources of a-priori knowledge, and the techniques used to embed such a knowledge base into real-world applications. Lastly, we end with the main conclusions and possible, future avenues for research.

2. Related work

HSE requires intermediate models of human motion to associate geometric knowledge with conceptual statements. Each type of representation will depend on the particular task to be accomplished. In our case, HSE demands, at the very least, the detection and tracking of relevant motion, the analysis of trajectory patterns, the interpretation of modeled behaviours, and the explanation of interpretations using conceptual terms.

Once the image sequences have been acquired from camera sensors, the first task is the detection of agents within the scene [5,31]. Several segmentation problems should be handled, such as illumination changes, shadows, camouflage, background in motion, and objects deposited into or removed from the scene [20]. These difficulties are coped by means of background modeling to determine foreground regions [25]. Among these methods, statistical approaches are very popular: W^4 uses a bimodal distribution [15]; Pfinder uses a single Gaussian per pixel to model the background [44]; Stauffer and Grimson use a mixture of Gaussians for modelling each pixel of the background model [40]; and Elgammal et al. present a non-parametric background model instead [10]. Common for these methods is that clutter in the background will therefore have an effect on the results.

Additionally, tracking procedures are incorporated to maintain target identification and to recover from segmentation errors, mainly due to drastic illumination changes and occlusions [39]. On the one hand, since tracking requires reasoning over time under uncertainty, a probabilistic framework is commonly used [6]. Specifically, several authors have focused in on tracking by means of Particle Filtering (PF): Nummiaro et al. [29] use a PF based on colour-histogram cues. However, no multiple-target tracking is considered. Pérez et al. [30] propose also a PF based on a colour-histogram likelihood. No appearance model updating is performed, which frequently leads to target loss in dynamic scenes. On the other hand, deterministic approaches have been introduced recently for rigid, single target tracking [9,8]. In particular, Comaniciu et al. [9] developed a tracking technique called *mean-shift*. However, their method tracks just one target, initialised by hand, and the appearance model is never updated. Collins et al. [8] presented an effective tracker, based also on the mean-shift algorithm, with on-line selection of discriminative features. However, this approach may suffer from model drift.

Once the agent model is properly tracked over time, it is possible to generate high-level descriptions. For this purpose, some a-priori knowledge about the environment, within which the behaviour is being observed, is included. However, interpretation should lead with the incompleteness and noise in the agent state due to detection and tracking errors. In order to cope with this uncertainty, interpretation may be learnt using a probabilistic framework, such as Mixtures of Gaussians (MoG) [26] or Belief Networks (BN) [17,34]. Alternatively, the temporal and uncertainty aspects of interpretation can be handled using logical schemes in a goal-oriented manner, that is, explicitly and hierarchically represented, not coded into conditional probabilities [14]. As example, Fuzzy Metric Temporal Horn Logic (FMTHL) can cope with these aspects of interpretation [38]: logic productions evolve over time as the received data does, so conceptual knowledge is time-delimited, and thus the development of events can be comprehended and even anticipated.

Due to the practical impossibility of modeling all possible human behaviours, the expected evolution of logical schemes is modeled a-priori for improving spatio-temporal interpretation [27]. Thus, the set of logic predicates to be instantiated are classified according to different criteria, such as specialization relationship [21], semantic nature [34] or temporal ordering [17]. This classification organizes the constituent semantic predicates of a human behaviour model into hierarchical structures, such as networks [37] or trees [24]. Towards this end, Situation Graph Trees (SGTs) has been proven suitable to represent the specialization, semantic and temporal relationships of a particular behaviour [14].

Finally, natural-language sentences are built upon the instantiated primitives of the behaviour model [7,22]. In smart video surveillance, human behaviour can be represented by scenarios, i.e. predefined sequences of events [41]. The scenario is evaluated and automatically translated into NL texts by analyzing the contents of the image sequence over time, and deciding on the most suitable predefined event that applies in each case. Obeying these

requirements, Discourse Representation Theory (DRT) is of particular interest, since it discusses algorithms for the translation of coherent NL texts into computer-internal representations based on logic predicates [12]. Thus, DRT assures syntactical correctness for the generated texts by means of four linguistic tasks, namely lexicalization, text generation, morphological changes, and orthography.

3. An architecture for human sequence evaluation

In order to combine the aforementioned tasks, HSE is here implemented based on a mixture of multiple co-operating top-down and bottom-up processes that make use of a-priori knowledge, see Fig. 1 [19,23,27]. So evaluation of human behaviour is a transformation process between pixel values and NL texts.

The first level is called *Active Sensor Level* (ASL). This layer acquires raw video data and information about camera parameters. Thus, it might be possible to control the viewing conditions, such as the viewpoint and the image resolution. Pixel values are forwarded to the *Image Signal Level* (ISL), where the sequence of image data is processed for human motion detection. The resulting moving or foreground regions are the basis for the following layer called *Picture Domain Level* (PDL). Possible segmentation errors generated at the ISL are handled here by means of tracking techniques. At the *Scene Domain Level* (SDL), the 3D configuration of the scene is used to compute the parameters of each agent within its 3D environment.

Tracking results obtained at either PDL or SDL are forwarded to the *Conceptual Integration Level* (CIL) to instantiate semantic predicates for a given agent and time step. These qualitative descriptions also become conceptual relationships of such an agent w.r.t. its environment. Instantiated predicates are handled at the *Behaviour Interpretation Level* (BIL) where the expected temporal evolution of descriptions are modeled a-priori to generate coherent spatio-temporal interpretations.

Interpretation results are handled at the *User Interaction Level* (UIL) for NL text generation. Broadly speaking, logical predicates generated at the BIL are associated to proper lemmata obtained from an a-priori corpus. Then, text generation rules are specified to (i) infer the syntactical order of the lemmata and (ii) inflect proper morphological cases.

Summarizing, HSE embeds three basic tasks. Firstly, the estimation of spatio-temporal descriptions of human motion in terms of numerical knowledge (ASL, ISL, PDL, and SDL). Secondly, the association of geometric parameters with semantic predicates (CIL and BIL). Thirdly, the generation of conceptual descriptions about the meaning of human motion, which is *explained* using NL texts (UIL). Next, a concrete implementation of this architecture is illustrated.

4. A knowledge base for HSE

Given a modular description of an HSE system, the sources of knowledge are next identified by showing the main characteristics of a real-world HSE application in an outdoor scene: two concrete example implementations will bring to light the requirements of generic HSE systems. In particular, a pedestrian crossing and a cafeteria are chosen as suitable experimental scenes, where multiple agents exhibiting different behaviours have been recorded, see Fig. 2(a).

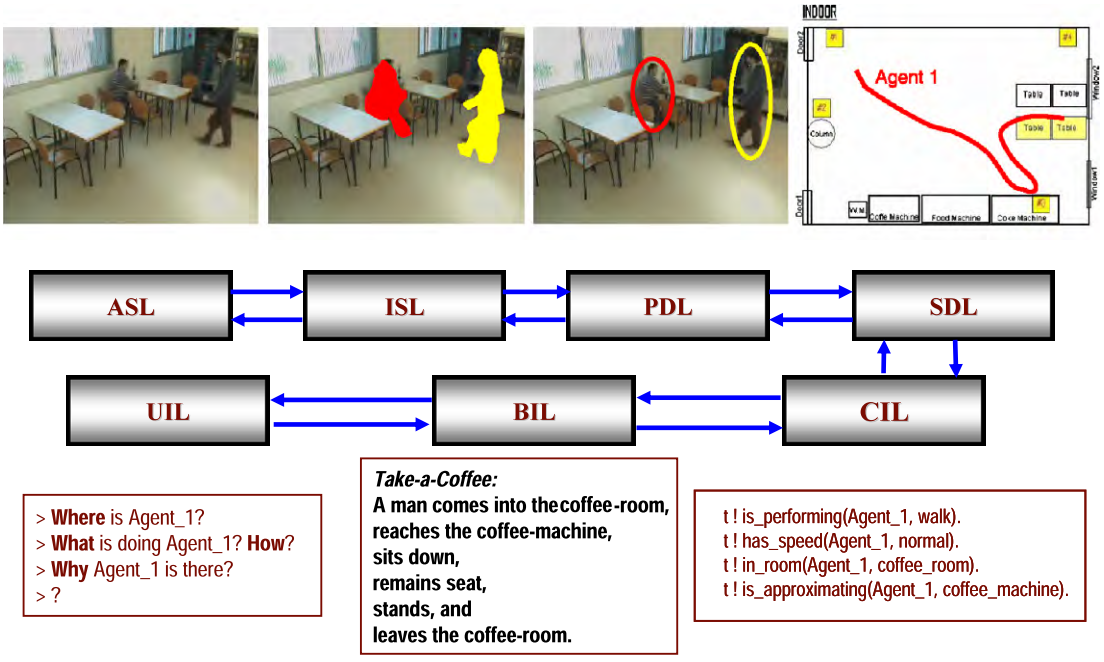

Fig. 1. A modular architecture for Human Sequence Evaluation in a human-populated scene. Further explanation of these levels is found in Section 3.

Fig. 2. (a) Two examples of image sequences, a *crosswalk* and a *cafeteria*, where a semantic interpretation of human behaviour can be accomplished. (b) Segmented pixels are painted in white, which correspond to foreground, dark and light foreground pixels. (c) Detection example: white ellipses represent each target, and purple lines denote their contour.

4.1. The active sensor level

This level feeds the system with image data. When information about camera parameters is provided, it is possible to represent the viewing conditions, such as the recording technique, the viewpoint and the image resolution. Moreover, the scene could be recorded using active sensors, thus allowing to vary the recording conditions. Thus, when an agent is detected far away in a scene, zooming in on the agent would provide the system with more detailed images. In this case, an image mosaic could be built, and the entire process would be transparent for the architecture here presented.

4.2. The image signal level

Once the image data is available, the system detects foreground objects which correspond to moving pixels defined w.r.t a background model. The segmentation process uses either colour or intensity cues to build the background model according to the sensor response, and the background is modelled on a pixel-wise basis. This is carried out during a training period by using a window of N frames, during which a motion filter is used to remove moving pixels. During such a training period, a Background Colour Model (BCM) is computed by calculating the mean and standard deviation for each colour channel of every image pixel. Based on the BCM, pixels are classified as foreground, dark and light foreground, highlight, shadow and background, see [16] for details. However, the BCM is not built for those pixels which are beyond the sensor linear range, i.e. they lack of colour or are over a saturation intensity value. In these cases, a Background Intensity Model (BIM) is built, which consists of the mean pixel intensity and its standard deviation. Thus, those pixels without BCM are segmented using a mean filter based on the BIM. Examples of segmentation are show in Fig. 2(b).

Once the current image has been segmented into the aforementioned five categories, blobs that may correspond to agents are detected. First, foreground, dark and light foreground maps are fused. Then, majority, opening and closing morphological operations are applied. Finally, a minimum-area filter is used and the surviving pixels are grouped into blobs. Subsequently, blobs are parametrically represented using an orientable ellipse – keeping the blob first and second order moments – so that spurious structural changes that blobs may undergo are constrained [29,9]. Thus, the j-observed blob at time t is given by:

$$\mathbf{z}_t^j = (\tilde{x}_t^j, \tilde{y}_t^j, \tilde{h}_t^j, \tilde{w}_t^j, \tilde{\theta}_t^j)^T \tag{1}$$

where $\tilde{x}_t^j, \tilde{y}_t^j$ represents the ellipse centroid, $\tilde{h}_t^j, \tilde{w}_t^j$ are the major and minor axes, respectively, and the $\tilde{\theta}_t^j$ gives the angle between the abscissa axis and the ellipse major one. Fig. 2(c) shows examples of target detection.

4.3. The picture domain level

Tracking at the PDL aims to establish coherent relations of the different targets between frames. This implies the inference of the state of each target within the scene using all evidence up to date. In this paper, the observation vector at time t for the agent j is given by the target state:

$$\mathbf{x}_t^j = (x_t^j, \dot{x}_t^j, y_t^j, \dot{y}_t^j, h_t^j, \dot{h}_t^j, w_t^j, \dot{w}_t^j, \theta_t^j)^T, \tag{2}$$

which defines a state variable for every observation one and adds the target speed and the size change rate. The angle θ_t^j is considered here to undergo minor variations, i.e. humans will remain upright in the sequence.

Here, the inference of $\mathbf{x}_t^j$ is the result of the conjunction of detection, estimation and adaptation tasks. These processes are addressed by splitting the tracking task into two types: short-term blob tracking – or Low-Level Tracking (LLT) – and long-term target tracking – or High-Level

Tracking (HLT). The former performs a motion-based tracking where detailed models are avoided. No appearance information is used, and events are not analysed. The latter builds appearance-based models and selects the most appropriate tracking approach according to these. The tracking results are shown in Fig. 3.

4.3.1. Low-level tracking

Low-level trackers establish coherent target relations between frames by setting correspondences between observations and trackers, and by estimating new target states according to the associated observations. We implemented each LLT as a linear filter, but the inclusion of more complex filters is straightforward. To start with, the Kalman filter [4] implements a recursive algorithm which works in a prediction–correction way, estimating the system state from noisy measures.

In a multiple-target tracking scenario, different observations could have been generated by clutter or noise, or might correspond to the same target. Consequently, we compute the regions where target observations are expected, in agreement with the target state and the system uncertainties. These regions, commonly denoted as *gates*, are computed based on the innovation covariance matrix which defines an ellipsoid in the observation space. Then, measures are associated to the nearest-neighbour tracker in whose gate they lie, and a bank of Kalman filters is used to estimate the state of all detected targets.

4.3.2. High-level tracking

Since the PDL includes appearance-based trackers, these are used to overcome motion-based tracking problems such as grouped targets, occlusions, camouflages, and segmentation errors. In these cases, HLT is used to obtain state estimates for each target, see [35] for details.

Two sources of information are used to decide which pixels are considered as belonging to each target: the silhouette of the associated observation given by the ISL and the filtered ellipse given by the LLT. The intersection is cropped for each target and its appearance is represented by means of colour-histograms [8,9,29].

Subsequently, a distracter-robust mean-shift technique is used. This achieves target localisation by performing a gradient-descent search on a image region of interest, which is previously weighted. The similarity between two histograms is computed using the Bhattacharyya metric. Thus, the mean-shift procedure recursively moves the candidate position to a new location, while searching the local maximum according to the aforementioned metric [9].

4.3.3. Management of LLT and HLT

A management process performs the association between Motion- and Appearance-based Tracking. The process consists of three cases [36]: firstly, once a LLT is confirmed, a HLT is instantiated and associated. In case that the new-born HLT corresponds to an isolated target, the target appearance is then computed. Secondly, if a LLT is already associated to a HLT, its parameters relative to the target position and shape are updated in subsequent tracker matchings. Thirdly, if LLTs have been removed when target segmentation is not feasible, those HLT with no correspondence with any LLT are tracked in a top-down process using appearance information. When a target is again detected after occlusion or grouping, a new LLT is instantiated. If this track becomes stable, it is confirmed and a new HLT is created. Then, the association process is performed by using the similarity between histograms of the lost HLT and the new one. Thus, the system may conclude that both trackers are in fact representing the same target.

The relations of LLT/HLT and the HSE framework are shown in Fig. 4. Thus, HLT acts on the lower levels following a top-down approach in numerous ways: by selecting motion-based or appearance-based tracking mode, by

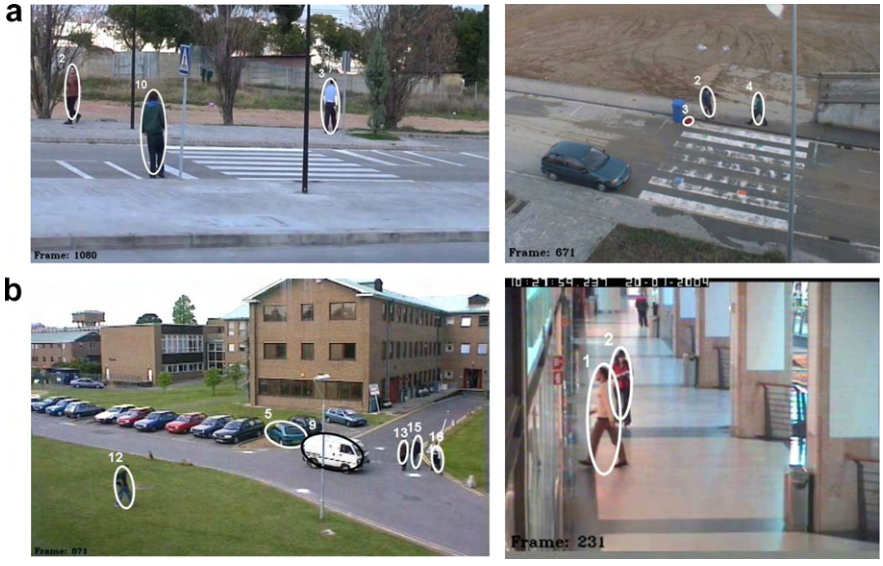

Fig. 3. Sample results of the tracking process: (a) shows results from Low-Level Tracking, while (b) shows results from High-Level Tracking.

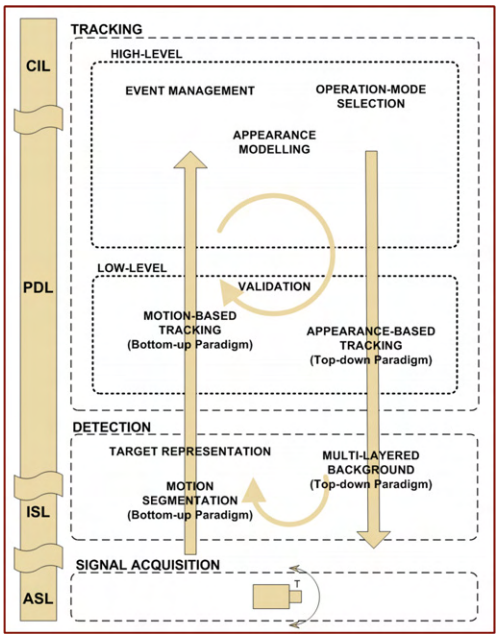

Fig. 4. Relations between the HSE architecture and the tracking tasks. Since we do not perform 3D tracking but only calibration, the PDL feeds directly to the CIL.

preventing the creation of nonfeasible LLTs, by validating the association of observations to LLT, by associating several LLTs to the same HLT, and by enabling the incorporation of a motionless objects into the background. Cognitive levels of HSE require the target state of all detected objects within the scene, as described next.

4.4. The scene domain level

At the SDL, models and algorithms are based on 3D configurations. We restrict ourselves to perform a calibration process for obtaining the 3D position estimation of the agents within the scene, based on their 2D coordinates in the image, see Fig. 5(a). As a result, we can estimate the real sizes of the agents and the 3D relationships between these agents and predefined objects within the scene, see Fig. 5(b). Thus, 3D knowledge of the scene helps to cope with occlusions due to other static components and also to grouping.

4.5. The conceptual integration level

All the conceptual knowledge used for HSE is implemented at the CIL as a set of logic predicates in FMTHL. Thus, we cope in a goal-oriented manner with the temporal and uncertainty aspects due to the integration of quantitative values into conceptual terms. This predicate logic language treats in a unified manner dynamic occurrences, uncertainties of state estimation, and intrinsic vagueness of conceptual terms [38].

On the one hand, the state vector of the agent $\mathbf{x}_t^j$ embeds quantitative knowledge related to dynamical, positional and postural properties of a given agent j at time t:

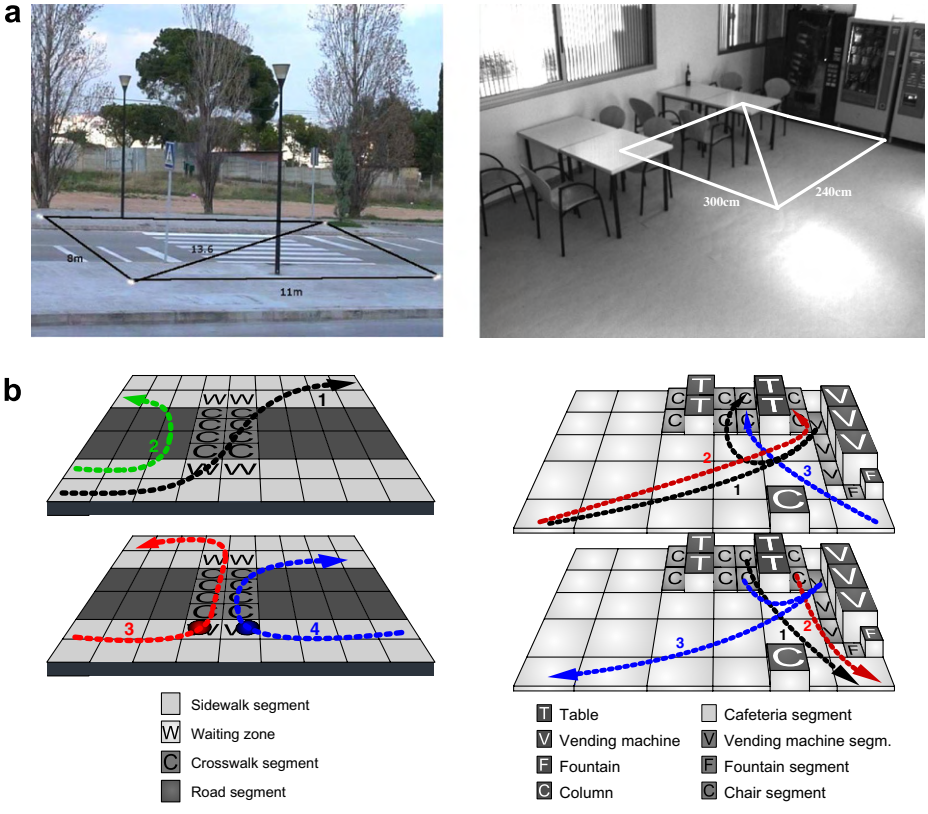

Fig. 5. (a) Calibration procedure and (b) the schematic trajectories of each agent in the floor plane for the *crosswalk* and *cafeteria* scenes.

$$\mathbf{s}_t^j = has\_status(Agent\_j, X, Y, Vel, Or, aLabel), \quad (3)$$

thus embedding the 2D spatial position (X, Y) in the floor plane, in addition to the velocity and orientation computed from $\mathbf{x}_t^j$ after calibration. These quantitative parameters are associated to semantic concepts like *moving*, *small*, *left*, or *briefly* with a fuzzy degree of validity characterizing how good a concept matches the numerical parameter value. For example, fuzzy attributes for speed and orientation instantiate predicates such as *has_speed* or *has_direction*, respectively. The action parameter *aLabel* distinguishes between three different actions based on the velocity, namely *running*, *walking* and *standing*, thus instantiating predicates such as *is_moving* or *is_waiting*. This is done by applying the trapezoidal membership function of Fig. 6.

On the other hand, the spatial relations of each agent w.r.t. its environment are derived. This is implemented by applying a distance function between the positions of the different agents/objects in the scene. Subsequently, a discretization of the resulting distance value is obtained by using fuzzy logic, thus allowing to instantiate predicates, such as *far*, *is_alone* or *has_distance_to*. Other spatial relationships are derived from the semantics of the environment, so a conceptual scene model is used to identify semantic locations. Thus, the scene model is divided into polygonally bounded segments which describe the possible positions in which an agent can be found, see Fig. 7.

Thus, each segment has an associated conceptual description. For the *crosswalk* sequence, we distinguish (at least) four different types of segments, namely: *sidewalk_segment*, *waiting_line*, *road_segment*, and *crosswalk*. As a result, we can then instantiate predicates which relate the spatial position of the agent w.r.t these segments, see Fig. 8. In this example, if the position X, Y of the *Agent_4* in the *crosswalk* scene lies within the limits of *segment_31* which is a *road_segment*, then the predicate *not_on_crosswalk* is said to be instantiated.

4.6. The behaviour interpretation level

All the aforementioned conceptual knowledge about $\mathbf{s}_t^j$ generated at the CIL is called a *situation* in [27]. Due to

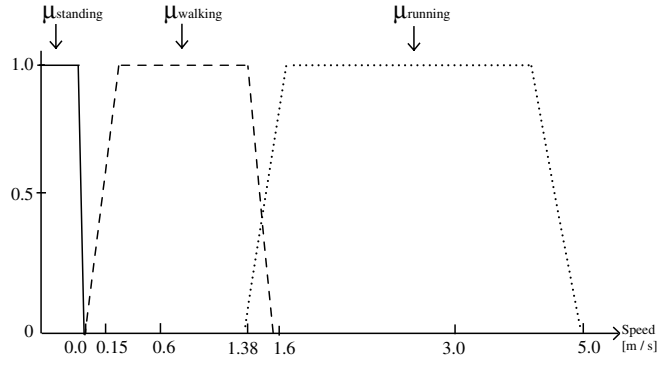

Fig. 6. Deriving the action based on a trapezoidal membership function.

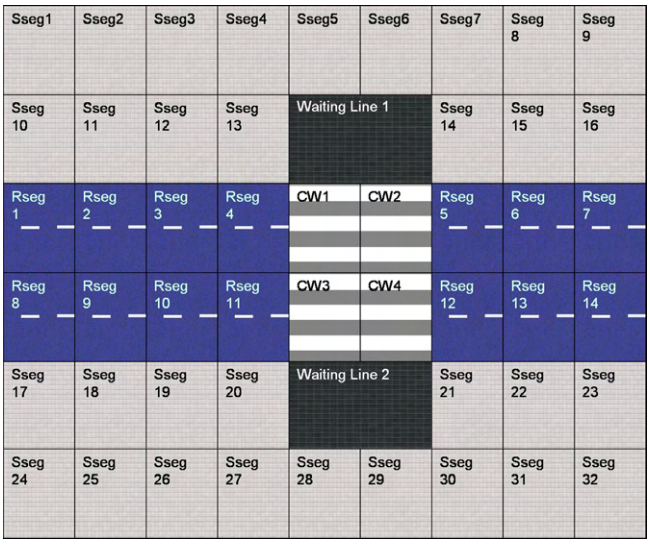

Fig. 7. Conceptual scene models for the *Crosswalk* sequence.

the impossibility of modeling all possible human situations, the expected evolution of situations to be described are modeled a-priori for improving spatio-temporal interpretation. That means, the BIL selects those situations to be instantiated, thus allowing to interpret the intentions of the agent in a goal-oriented manner. Towards this end, Situation Graph Trees (SGTs) constitute a suitable behaviour model which explicitly represents and combines the specialization, temporal, and semantic relationships of its constituent conceptual predicates, see Fig. 9 [14].

The basic component of SGTs is the *situation scheme*, which embeds the semantic predicates for a given agent at each frame step. Situation schemes embed two parts: the *state scheme* which refers to the conceptual knowledge which should be satisfied for instantiating the situation, and the *reaction scheme* which embeds the action supposedly carried out by the agent in that state. SGTs are then tree-like structures of Situation Graphs (SGs) which organize situation schemes into temporal sequences of other schemes using *prediction* edges.

In the example of Fig. 9, *LOOPING* is the start and *OUT OF BOUNDS* the end situations. Each path from a start to an end situation defines a sequence of situations represented by the SG. Also, SGs can particularize superordinate situations by using *particularization* edges. These edges allow to describe a situation into a conceptually or temporally more detailed manner. In the example, the situation *AGENT ACTIVE* is the parent of the specialized situation *LOOPING*.

SGTs can be used to recognize those situations which can be instantiated for an observed agent by applying the so called *graph traversal* [14]. The goal is to determine the most particularized situation which can be instantiated by considering the FMTHL predicates that are true at each time step, according to $\mathbf{s}_t^j$. This traversal of the SGT is applied by considering the knowledge encoded in the form of prediction and particularization edges and is imple-

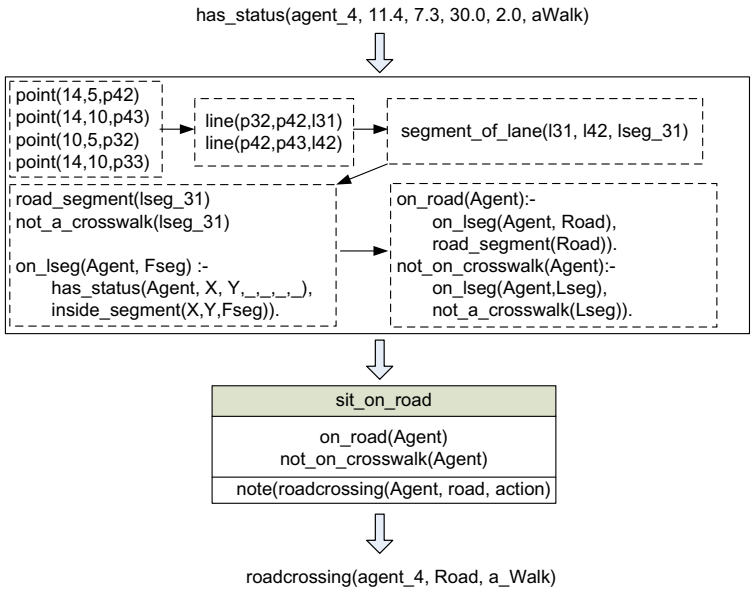

Fig. 8. From quantitative knowledge to conceptual predicates.

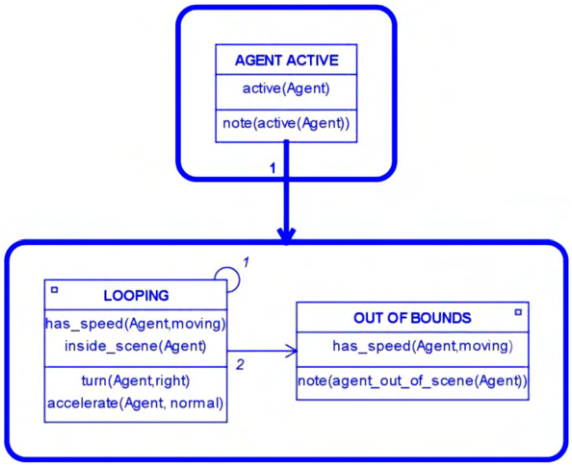

Fig. 9. Example of Situation Graph Tree.

mented by executing a logic-program obtained by an automatic translation of the SGT into rules of FMTHL, see [3] for details.

Thus, bottom-up data flow corresponds to potential semantic descriptions (hypotheses based on estimations) derived from motion analysis processes. In particular, the reaction predicate *note* produce textual concepts to be reported to end-users, as reflected in Fig. 10 where two SGTs are used to interpret the *crosswalk* and the *cafeteria* sequences.

In Fig. 10(a), three different behaviours are modeled: (i) a pedestrian walks on the sideway and crosses the pedestrian crossing without stopping to see whether a car is approaching; (ii) a pedestrian stops on the waiting line for a few seconds before continuing; and (iii) a pedestrian crosses the road neither by using the pedestrian crossing nor stopping. Table 1 shows the conceptual descriptions generated for *agent_3*.

Also, three different behaviours are modeled by the SGT depicted in Fig. 10(b) for the *cafeteria* scene: (i) a person walks by the cafeteria; (ii) a person sits on a chair; and (iii) a person waits in front of a vending machine. Table 2 shows the conceptual descriptions generated for *agent_1*.

All these predicates are forwarded to the UIL to generate NL texts, as described next.

4.7. The user interaction level

The first step involved at UIL is the elaboration of a *corpus* made by native speakers. In our case, six different people have contributed to the corpus with their own descriptions of the *crosswalk* and *cafeteria* scenes in Catalan, Spanish and English languages. Once the corpus has been built, the grammatical categories (noun, adjective, verb, ldots) are established for all the lemmata.

Then, conceptual information is transformed into linguistic outputs using Discourse Representation Theory (DRT) [18,2]. Thus, the content and structure for the output are determined, including the ordering of the different pieces of information that will be transformed into NL texts. In our work, Catalan and Spanish languages have been included in a DRT-based system called `Angus2` [12]. Towards this end, the addition of a new language using DRT involves four linguistic tasks [32]: lexicalization, text generation rules, morphological rules and changes in the orthography, see Fig. 11.

At the lexicalization step, the conceptual predicates are clustered into appropriated lemmata by means of an ordered set of language-dependent rules. After that, Text Generation Rules (TGRs) are specified for each language to infer the syntactical order of the input lemmata. The TGR step decides also those expressions which best identify the entities of a particular domain. Thus, the symbolic

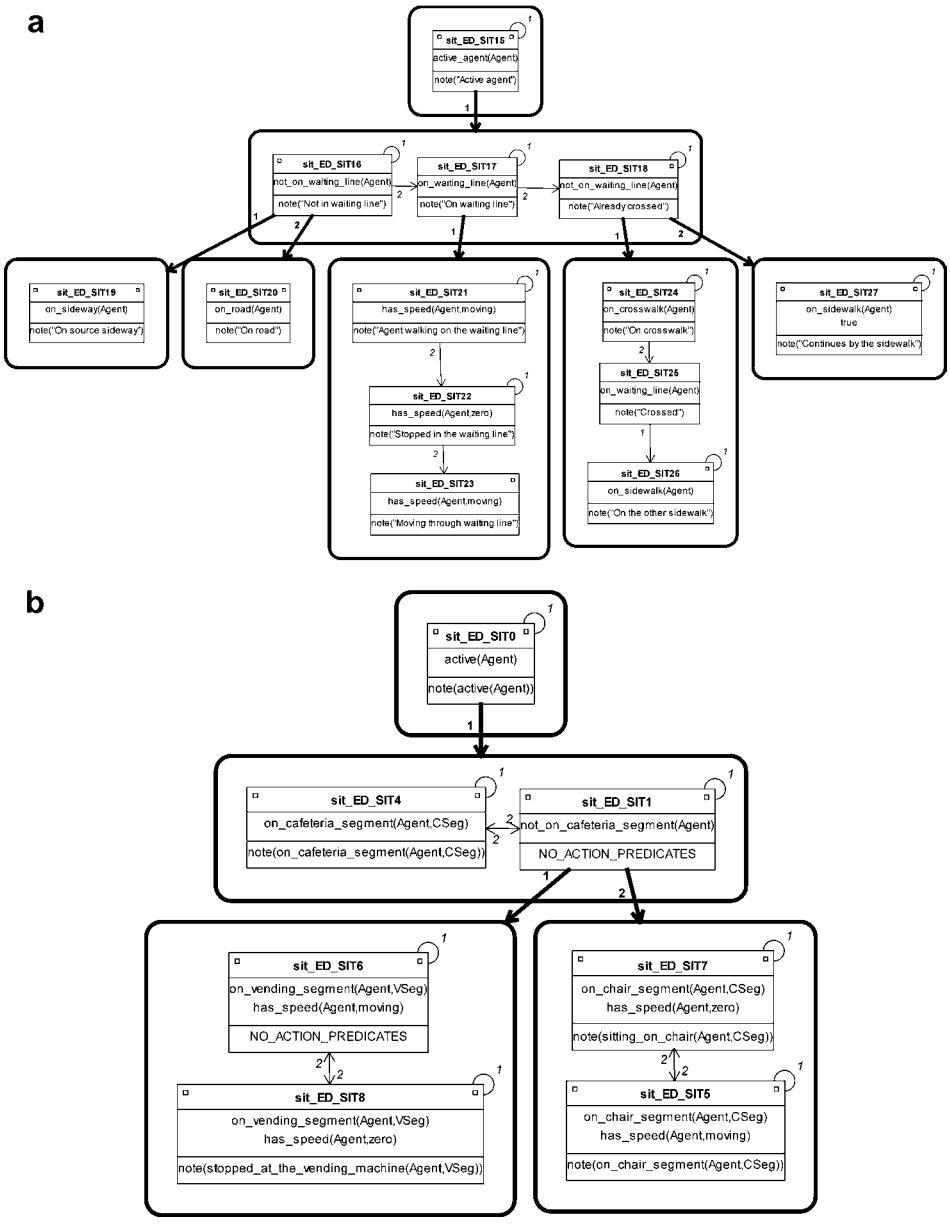

Fig. 10. SGTs used in (a) the *Crosswalk* and (b) the *Cafeteria* sequences.

Table 1
Logic predicates for *Agent*_3 at the *crosswalk*

Start	Situation
203	on_source_sideway(**agent_3**,sseg_32)
230	on_source_sideway(**agent_3**,sseg_31)
252	on_source_sideway(**agent_3**,sseg_30)
288	on_source_sideway(**agent_3**,sseg_29)
326	agent_walking_on_the_waiting_line(**agent_3**)
401	stopped_in_the_waiting_line(**agent_3**)
506	moving_through_waiting_line(**agent_3**)
512	on_crosswalk(**agent_3**)
616	crossed(**agent_3**)
669	on_the_other_sidewalk(**agent_3**)

names of the knowledge base entities are replaced by the semantic content of textual components [32]. Subsequently, morphological rules are applied over the set of lemmata to

Table 2
Logic predicates for *Agent*_1 at the *cafeteria*

Start	Situation
205	on_cafeteria_segment(**agent_1**,cseg_11)
221	on_cafeteria_segment(**agent_1**,cseg_8)
237	on_cafeteria_segment(**agent_1**,cseg_9)
279	on_vending_segment(**agent_1**,vseg_3)
289	stopped_at_vending_mach(**agent_1**,vseg_3)
337	on_vending_segment(**agent_1**,vseg_3)
342	on_cafeteria_segment(**agent_1**,cseg_6)
360	on_cafeteria_segment(**agent_1**,cseg_9)
377	on_chair_segment(**agent_1**,chseg_3)
412	on_chair_segment(**agent_1**,chseg_4)
431	sitting_on_chair(**agent_1**,chseg_4)
1409	on_chair_segment(**agent_1**,chseg_4)
1420	on_chair_segment(**agent_1**,chseg_3)
1455	on_cafeteria_segment(**agent_1**,cseg_9)
1472	on_cafeteria_segment(**agent_1**,cseg_6)

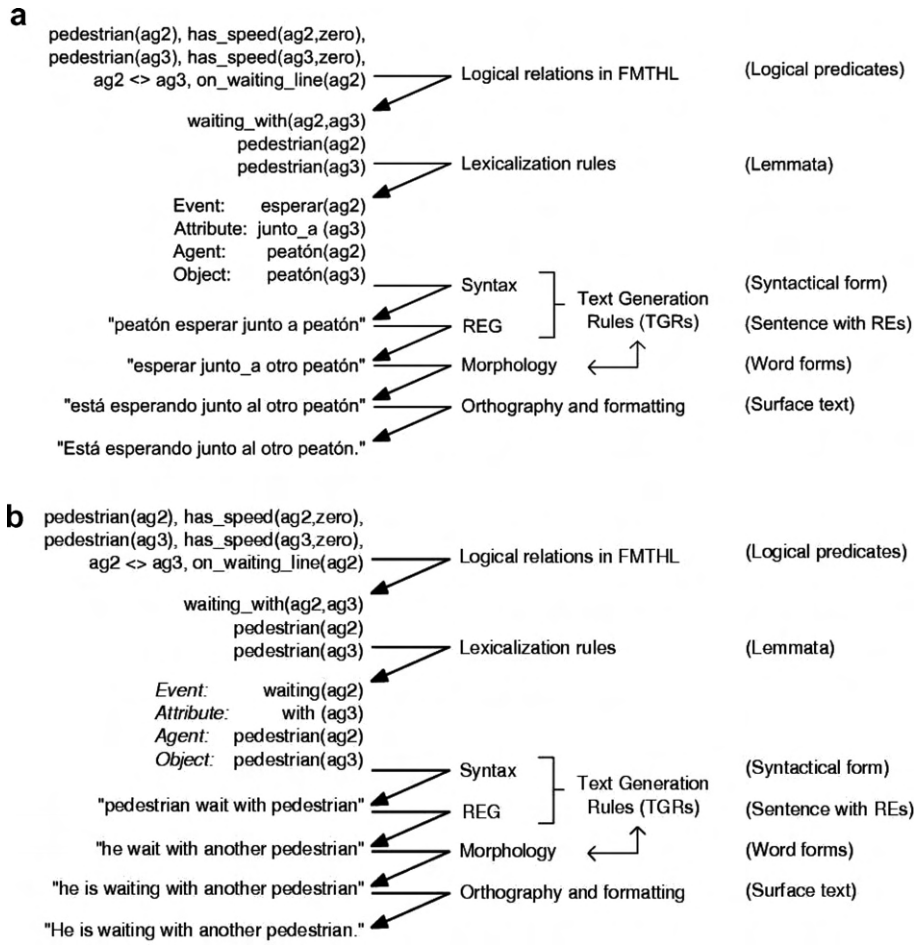

Fig. 11. Example for the generation of textual sentences in (a) Spanish and (b) English. The center column contains the tasks to be performed, and the left column indicates the outputs obtained.

properly inflect the linguistic elements (number, gender, tense,...). Lastly, orthography provides punctuation symbols to the sequence of words to be delivered to the final user.

As a result, given the trajectories depicted in Fig. 12, Table 3 summarizes the resulting textual descriptions embedding the *explanation* provided by the HSE system. Multi-lingual capabilities are demonstrated using Catalan, Spanish and English languages: sentences embed semantic knowledge about human behaviour, and the textual generation is linguistically correct.

5. Conclusions

This paper is focused on the transformation process of image data into semantic descriptions. This process, called Human Sequence Evaluation, evaluates the recorded human behaviour in image sequences for the automatic

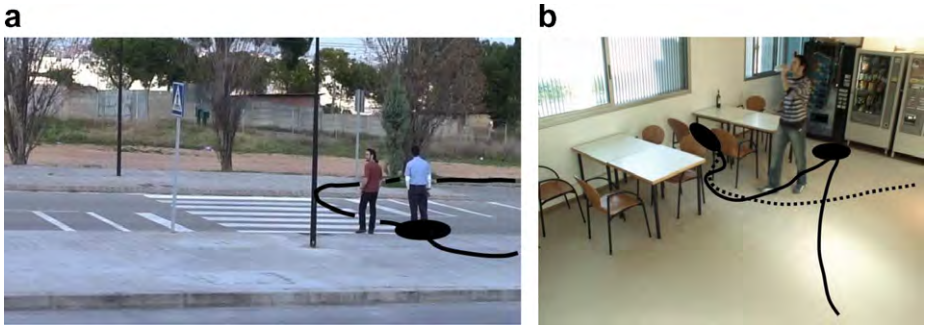

Fig. 12. Trajectories for (a) Agent_3 and (b) Agent_1 explained in Table 3.

Table 3
Resulting NL texts generated for *Agent_3* in the *crosswalk* scene and for *Agent_1* in the *cafeteria* scene

Catalan and English description for *Agent_3*		
203	*Lo vianant surt per la part inferior dreta*	*The pedestrian appears by the lower right side*
252	*Va per la vorera inferior*	*He walks by the lower sideway*
401	*S'espera per creuar*	*He waits to cross*
436	*S'está esperant amb un altre vianant*	*He is waiting with another pedestrian*
506	*Creua pel pas zebra*	*He crosses by the crosswalk*
616	*Va per la vorera superior*	*He walks by the upper sidewalk*
749	*Se'n va per la part superior dreta*	*He leaves by the upper right side*
Spanish and English descriptions for *Agent_1*		
205	*La persona aparece por la parte inferior dreta*	*The person appears by the lower right side*
237	*Camina por la cafetería*	*He walks by the cafeteria*
289	*Se espera en la máquina de ventas*	*He waits at the vending machine*
342	*Camina por la cafetería*	*He walks by the cafeteria*
377	*Camina entre las sillas*	*He walks among chairs*
431	*Se sienta en una silla*	*He sits on a chair*
1409	*Camina entre las sillas*	*He walks among chairs*
1455	*Camina por la cafetería*	*He walks by the cafeteria*
1480	*Se va por la parte superior derecha*	*He leaves by the upper right side*

The trajectories of both agents are shown in Fig. 12.

generation of natural-language text descriptions in different languages.

In essence, HSE (i) segments agent motion – i.e. trajectories – into individual movements, (ii) treats the relation between the movements implicitly by the order in which these movements are described, and (iii) then converts the result into coherent NL paragraphs. These main steps are organized within an architecture based on a set of co-operating modules, each one devoted to a specific task: the estimation of spatio-temporal descriptions of human motion in terms of quantitative knowledge (ASL, ISL, PDL and SDL); the association of geometric parameters with semantic predicates (CIL and BIL); and the generation of NL texts explaining the meaning of observed human motion (UIL).

Subsequently, concrete example implementations of this architecture is presented for different scenes, i.e. a pedestrian crossing and a cafeteria: a system that evaluates video sequences about human behaviours by generating natural-language descriptions in Catalan, Spanish and English has been successfully developed in a first stage.

Further steps include the enhancement of current tracking capabilities by including the detection of groups and interactions between agents. Research on artificial cognition will enhance the behaviour model to manipulate *networks of meanings*. Additionally, HSE might incorporate object recognition and automatic learning capabilities to increase the corpus of the resulting texts. Subsequently, these texts can be used for automatic video annotation, i.e. to provide content description in textual form about human behaviours observed in videos.

Acknowledgements

The authors would like to thank Prof. Hans-Hellmut Nagel for his suggestions, corrections and remarks in this paper and towards Image Sequence Evaluation. Also, we wish to thank Pau Baiget, Carles Fernández, Ivan Huerta, and Tere Izquierdo for their valuable help in this work. Thanks to the anonymous referees for their comments which inspired us to enhance the initial version.

References

[1] J.K. Aggarwal, Q. Cai, Human motion analysis: a review, Computer Vision and Image Understanding 73 (3) (1999) 428–440.
[2] M. Arens. SGTEditor v1.0 Reference Manual v1.1. IAKS, Fakultät für Informatik der Universität Karlsruhe (TH), April 2003.
[3] M. Arens, A. Ottlik, H.-H. Nagel, Natural language texts for a cognitive vision system, in: ECAI, Lyon, France, 2002, pp. 455–459.
[4] Y. Bar-Shalom, T. Fortmann, Tracking and Data Association, Academic Press, 1988.
[5] A.F. Bobick, J. Davis, The representation and recognition of movement using temporal templates, TPAMI 23 (3) (2001) 257–267.
[6] D.J. Bullock, J.S. Zelek, Real-time tracking for visual interface applications in cluttered and occluding situations, Image and Vision Computing 22 (12) (2004) 1083–1091.
[7] H. Buxton, Learning and understanding dynamic scene activity: a review, Image and Vision Computing 21 (1) (2002) 125–136.
[8] R. Collins, Y. Liu, M. Leordeanu, Online selection of discriminative tracking features, TPAMI 27 (10) (2005) 1631–1643.
[9] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, TPAMI 25 (5) (2003) 564–577.
[10] A. Elgammal, D. Harwood, L.S. Davis, Nonparametric background model for background subtraction, in: ECCV, Dublin, Ireland, 2000, pp.751–767.
[11] D.M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding 73 (1) (1999) 82–98.
[12] R. Gerber, H.-H. Nagel, (Mis-?) using DRT for generation of natural language text from image sequences, in: ECCV, Freiburg, Germany, 1998, pp.255–270.
[13] J. Gonzàlez, Human Sequence Evaluation: the Key-frame Approach, Ph.D. thesis, UAB, Spain, 2004. Available from: <http://iselab.cvc.uab.es/>.
[14] M. Haag, H.-H. Nagel, Incremental recognition of traffic situations from video image sequences, Image and Vision Computing 18 (2) (2000) 137–153.

[15] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, TPAMI 22 (8) (2000) 809–830.
[16] I. Huerta, D. Rowe, M. Mozerov, J. Gonzàlez, Improving background subtraction based on a casuistry of colour-motion segmentation problems, in: ibPRIA, Girona, Spain, 2007, pp. 475–482.
[17] S.S. Intille, A.F. Bobick, Recognized planned, multiperson action, International Journal of Computer Vision 81 (3) (2001) 414–445.
[18] H. Kamp, U. Reyle, From Discourse to Logic, Kluwer Academic Publishers, Dordrecht Boston London, 1993.
[19] T. Kanade, Region segmentation: signal vs. semantics, in: Fourth International Joint Conference on Pattern Recognition, November 1978, Kyoto, Japan, pp. 95–105.
[20] M. Karaman, L. Goldmann, D. Yu, T. Sikora, Comparison of static background segmentation methods, in: Visual Communications and Image Processing, Beijing, China, July 2005, pp.2140–2151.
[21] I.A. Karaulova, P.M. Hall, A.D. Marshall, Tracking people in three dimensions using a hierarchical model of dynamics, Image and Vision Computing 20 (2002) 691–700.
[22] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions, International Journal of Computer Vision 50 (2) (2002) 171–184.
[23] M.H. Ma, P. McKevitt, Interval relations in lexical semantics of verbs, Artificial Intelligence Review 21 (3–4) (2004) 293–316.
[24] T. Matsuyama, V. Hwang, SIGMA: a framework for image understanding – integration of bottom-up and top-down analysis. in: Nineth International Joint Conference on Artificial Intelligence, vol. 2, LA, USA, 1984, pp. 908–915.
[25] T.B. Moeslund, A. Hilton, V. Kruger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2006) 90–126 (November–December).
[26] R.J. Morris, D.C. Hogg, Statistical models of object interaction, International Journal of Computer Vision 37 (2) (2000) 209–215.
[27] H.-H. Nagel, From image sequences towards conceptual descriptions, Image and Vision Computing 6 (2) (1988) 59–74.
[28] H.-H. Nagel, Steps toward a cognitive vision system, AI Magazine, Cognitive Vision 25 (2) (2004) 31–50.
[29] K. Nummiaro, E. Koller-Meier, L. Van Gool, An adaptive color-based particle filter, Image and Vision Computing 21 (1) (2003) 99–110.
[30] P. Pérez, C. Hue, J. Vermaak, M. Gangnet, Color-based Probabilistic Tracking, in: ECCV, Copenhaguen, Denmark, 2002, pp. 661–675.
[31] F. Porikli, Achieving real-time object detection and tracking under extreme conditions, Journal of Real-time Image Processing 1 (2006) 33–40.
[32] E. Reiter, R. Dale, Building Natural Language Generation Systems, Cambridge University Press, Cambridge, UK, 2000.
[33] P. Remagnino, G.L. Foresti, Ambient intelligence: a new multidisciplinary paradigm, IEEE Transactions on Systems, Man, Cybernetics-Part A: Systems and Humans 35 (1) (2005) 1–6.
[34] P. Remagnino, T. Tan, K. Baker, Agent oriented annotation in model based visual surveillance, in: ICCV, Mumbai, India, 1998, pp. 857–862.
[35] D. Rowe, J. Gonzàlez, I. Huerta, J.J. Villanueva, On reasoning over tracking events, in: 15th SCIA, Aalborg, Denmark, 2007, pp. 502–511.
[36] D. Rowe, I. Reid, J. Gonzàlez, J. Villanueva, Unconstrained multiple-people tracking, in: 28th DAGM, Berlin, Germany, 2006, pp. 505–514.
[37] G. Sagerer, H. Niemann, Semantic Networks for Understanding Scenes, Plenum Press, New York, 1997.
[38] K. Schäfer, C. Brzoska, F-Limette fuzzy logic programming integrating metric temporal extensions, Journal of Symbolic Computation 22 (5–6) (1996) 725–727.
[39] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, R. Bolle, Appearance models for occlusion handling, Image and Vision Computing 24 (11) (2006) 1233–1243.
[40] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, TPAMI 22 (8) (2000) 747–757.
[41] M. Thonnat, N. Rota, Image understanding for visual surveillance applications, in: Proceedings of Third International Workshop on Cooperative Distributed Vision, Kyoto, Japan, November 1999, pp 51–82.
[42] J.K. Tsotsos, Motion understanding: task-directed attention and Representations that link perception with action, International Journal of Computer Vision 45 (3) (2001) 265–280.
[43] L. Wang, W. Hu, T. Tan, Recent developments in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.
[44] C.R. Wren, A. Azarbayejani, T. Darrell, A.P. Pentland, Pfinder: real-time tracking of the human body, TPAMI 19 (7) (1997) 780–785.